

Accuracy Assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region

Zhiliang Zhu, Limin Yang, Stephen V. Stehman, and Raymond L. Czaplewski

Abstract

The U.S. Geological Survey, in cooperation with other government and private organizations, is producing a conterminous U.S. land-cover map using Landsat Thematic Mapper 30-meter data for the Federal regions designated by the U.S. Environmental Protection Agency. Accuracy assessment is to be conducted for each Federal region to estimate overall and class-specific accuracies. In Region 2, consisting of New York and New Jersey, the accuracy assessment was completed for 15 land-cover and land-use classes, using interpreted 1:40,000-scale aerial photographs as reference data. The methodology used for Region 2 features a two-stage, geographically stratified approach, with a general sample of all classes (1,033 sample sites), and a separate sample for rare classes (294 sample sites). A confidence index was recorded for each land-cover interpretation on the 1:40,000-scale aerial photography. The estimated overall accuracy for Region 2 was 63 percent (standard error 1.4 percent) using all sample sites, and 75.2 percent (standard error 1.5 percent) using only reference sites with a high-confidence index. User's and producer's accuracies for the general sample and user's accuracy for the sample of rare classes, as well as variance for the estimated accuracy parameters, were also reported. Narrowly defined land-use classes and heterogeneous conditions of land cover are the major causes of misclassification errors. Recommendations for modifying the accuracy assessment methodology for use in the other nine Federal regions are provided.

Introduction

A conterminous U.S. land-cover map is being developed at the U.S. Geological Survey (USGS) EROS Data Center using Landsat Thematic Mapper (TM) 30-meter-resolution imagery as the baseline data. This regional land-cover mapping project is jointly conducted by USGS and the U.S. Environmental Protection Agency (EPA), with the central objective to provide a generalized and regionally consistent land-cover product for use in a

broad range of applications. Each of the ten EPA Federal regions is mapped independently. An EPA Federal region consists of two or more States, and the ten regions make up the conterminous United States. Aspects of the mapping effort, ranging from teams of analysts to classification techniques, are consistent within each region but can vary among the regions.

At the core of this mapping project is a 23-category land-cover map (Table 1) produced using 1991-93 TM data for two dates: vegetation leaf-on and leaf-off. The two dates selected are usually within 1 year of each other. After radiometric and geometric corrections were applied, scenes for each region were spectrally stitched to form an image mosaic for further processing and analysis. The classification system and mapping techniques have been described in detail in Vogelmann *et al.* (1998).

Accuracy assessment is an integral component of any mapping project based on remote sensing. As the USGS land-cover map for each Federal region is completed, thematic accuracy is assessed to measure general and categorical qualities of the data. Assessing accuracy for the USGS regional mapping project is a complex task, largely owing to the size of the study areas relative to the 30-meter spatial resolution of the TM data used. Virtually no suitable reference data from existing survey programs can be used consistently for all Federal regions. Collecting new reference data is extremely labor intensive and time consuming, so a carefully chosen sampling design is necessary in order to use available resources efficiently. Developing a practical and statistically sound sampling plan that can characterize the accuracy of common and rare classes of the map product in such a large area is the key to an effective accuracy assessment.

Region 2, the smallest EPA federal region in area (over 181 million 30-meter pixels), consists of only the States of New York and New Jersey. Of the 23 land-cover classes, 15 were found in the region (Table 1). Region 2 was among the first regions mapped and, consequently, served as the prototype area for developing methodology for the accuracy assessment. The Region 2 accuracy assessment was designed to satisfy the following objectives:

- Develop a practical methodology to collect reference data based on a probability sampling design and a well-defined response design protocol.

Z. Zhu is with U.S. Geological Survey EROS Data Center, Sioux Falls, SD 57198 (zhu@usgs.gov).

L. Yang is with Raytheon, an on-site contractor with the U.S. Geological Survey EROS Data Center, Sioux Falls, SD 57198 (lyang@edcmail.cr.usgs.gov).

S.V. Stehman is with the State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210 (systema@mailbox.syr.edu).

R.L. Czaplewski is with the U.S. Forest Service, Rocky Mountain Research Station, Ft. Collins, CO 80526 (czaplewski_ray@rmrs.fs.fed.us).

Photogrammetric Engineering & Remote Sensing
Vol. 66, No. 12, December 2000, pp. 1425-1435.

0099-1112/00/6612-1425\$3.00/0

© 2000 American Society for Photogrammetry
and Remote Sensing

TABLE 1. SAMPLE SITES FOR THE GENERAL AND RARE-CLASS DESIGNS LISTED BY LAND-COVER CLASSES. OF THE 23 USGS REGIONAL LAND-COVER CLASSES (VOGELMANN *ET AL.*, 1998), 15 ARE FOUND IN REGION 2 AND ARE NUMBERED FROM 1 TO 15 FOR THE PURPOSE OF THIS PAPER.

Map Class Name	Class Number	General Sample	Percent of the General Sample	Percent of Map Pixels	Rare Sample
Open water			7.84	13.67	
Low intensity residential			5.13	4.21	
High intensity residential			1.94	1.23	43
High intensity commercial built-up			0.87	1.08	47
Hay/pasture			10.45	9.20	
Row crop			12.88	12.24	
Urban grass			0.77	0.75	42
Needleleaf evergreen forest			5.91	5.46	
Mixed forest			14.52	16.41	
Broadleaf deciduous forest			35.82	31.88	
Woody wetland			2.52	2.91	
Emergent herbaceous wetland			1.06	0.72	47
Quarry/strip mine/gravel pit			0.10	0.13	41
Bare rock/sand/clay			0.10	0.05	35
Transitional barren			0.10	0.07	39
Total:			100	100	
Classes not in Region 2:					
Small grain crop					
Bare soil					
Deciduous shrubland					
Evergreen shrubland					
Mixed shrubland					
Planted/cultivated woody plantation					
Grassland					
Perennial ice/snow					

- Describe site-specific thematic accuracy for all of Region 2 as the target population.
- Estimate overall accuracy as well as category-specific accuracy (i.e., user's and producer's accuracies (Congalton, 1991)), and
- Document details of the protocol for future reference, and note areas needing improvement.

In this paper we will describe methods and results of the accuracy study for EPA Region 2. We will also discuss lessons learned and their implications for planning subsequent accuracy assessments for the rest of the Federal regions.

Methods

A typical land-cover accuracy assessment contains three distinctive and integral phases: response design, sampling design, and analysis and estimation (Stehman and Czaplewski, 1998). This breakdown provides a convenient way to consider assessment features separately in the three parallel phases. The response design refers to how reference data are collected, whereas the sampling design deals with choosing a sample plan that is appropriate for project goals. Analysis and estimation are concerned with calculating accuracy estimates, along with the standard errors of those estimates. For this project, the response design includes the protocol for collecting information to determine the true land cover at a sample location, as well as for assigning the reference land-cover label. The sampling design component focuses on which elements of the target population are actually selected and the reference classifications that are assigned.

Response Design

Reference Data

When the study area is large and collecting field data is impractical, the choice of reference data is often limited to existing ancillary data sets. In this study, we reviewed several existing national programs to determine if any of them could be used as the source of reference data. These programs included the

National Resource Inventory (NRI) of the U.S. Department of Agriculture (USDA) Natural Resource Conservation Service, the Forest Inventory and Analysis (FIA) of the USDA Forest Service, the U.S. Department of Commerce National Agricultural Statistics Service, the U.S. Department of the Interior GAP Program, and the National Aerial Photography Program (NAPP). The usefulness of these data sets, except for NAPP, is limited in two aspects: incomplete coverage of the target population and different land-cover classification systems. For example, NRI data do not cover Federal lands, and FIA data are limited to forest land only. The differences in land-cover classification systems also hinder a direct comparison using these data sets.

As a national program, NAPP is flown systematically at approximately 5-year intervals over the entire country. Either black-and-white or color-infrared aerial photographs are recorded at the scale of 1:40,000. Because NAPP covers the whole country, it provides an adequate source of reference data from which to design a suitable sampling plan. The NAPP photographs taken in the early 1990s generally coincide with the date of the TM data used for the classifications. Using NAPP aerial photographs may result in interpretation error in the reference data. But the effect of interpretation error can be mitigated by developing consistent, well-documented response design protocols.

NAPP products for large-area land-cover accuracy assessment include scanned and terrain-rectified photographs in digital form (digital ortho quadrangles, or DOQ) and hardcopy NAPP photographs in either print or transparency form. In this study, NAPP photographic prints were preferred because they are easy to use and have sufficient resolution for photointerpretation. Producing transparencies requires extra steps and has no photointerpretation advantage over prints for our objectives. Complete coverage in DOQ is not available, making it inappropriate for a regional accuracy assessment, and the lack of TM-DOQ coregistration has the potential to compound errors.

Additionally, stereo viewing can be used in interpretation, and it can be particularly useful for certain classes. However,

our Region 2 experience showed that, at the cost of the extra effort for stereo viewing, it did not offer a substantial advantage over single photographs.

Unit of Assessment

Thematic accuracy can be evaluated using a variety of spatial units, including pixel blocks (e.g., 3 by 3 pixels), individual pixels, and polygons (Stehman and Czaplewski, 1998). In this study, pixels were used as the unit of assessment—the same as the basic mapping unit in the final USGS map products (unfiltered and unsmoothed). Without accounting for any spatial effects (e.g., salt-and-pepper, misregistration effects), results of this accuracy assessment reflect both misclassification and potential, albeit unmeasured, geometric aggregation factors.

Photointerpretation Protocol

Sample points (pixels) were located by overlaying their coordinates on the TM spectral image on the screen. Sample coordinates generated from the sampling design (discussed in the next section) were “copy-and-pasted” to the image cursor location and visually transferred to NAPP photographic prints. Displaying TM spectral bands in red-green-blue combination for the purpose of locating sample points provided two advantages. First, viewing the spectral image rather than the classified map maintained the objectivity of the photointerpretation process. Second, finding the corresponding locations on the non-georeferenced NAPP prints was eased by visually consulting with spatial patterns (but not map classes) apparent on the TM color composite image.

Once sample coordinates were transferred from the TM image on the screen to the NAPP prints, the sample sites were interpreted directly on the photographs. Reference land-cover labels and attributes were visually interpreted and recorded onto a spreadsheet file. For each record, the following fields of information were obtained:

- Primary and secondary land cover of the sample site
- Dominant land cover of adjacent pixels
- Relative location of the sample site
 - (1) On the edge of two land-cover classes
 - (2) Homogeneous (one land-cover class)
 - (3) Heterogeneous (more than two land-cover classes)
- Confidence of photointerpretation
 - (1) Land-cover and land-use information is too difficult to interpret
 - (2) Interpretation is perhaps a correct label but there is some doubt
 - (3) Interpretation is probably a correct label
 - (4) Interpretation is absolutely a correct label
- Notes on other factors affecting the photointerpretation (e.g., temporal effects).

Photointerpretation does not always result in precise, unambiguous land-cover labels. Closely related land-cover types, such as conifer and mixed forests, are usually the cause of uncertainties in defining a correct classification. In such situations, both primary and secondary land-cover descriptions were recorded, and either of the two would be considered correct.

The primary motivation for recording confidence and relative location information during interpretation was to provide opportunities to address issues related to misclassification or photointerpretation at a later stage of analysis. Relative location informs us about mixed pixel problems, and the confidence information is related to uncertainties of photointerpretation. Additionally, possible land-cover changes resulting from differences in TM and NAPP dates are captured in both the confidence index and supplemental notes. Often in just 1 year, crop types may be changed (e.g., from row crops to hay and pasture) or forest lands cleared. These temporal changes, which

could be, and indeed were, interpreted on the basis of tonal differences from the TM imagery, were the result of the NAPP date being different from the TM date and should be differentiated from true misclassification error.

To minimize human errors in the photointerpretation process, at least two analysts examined the same set of sample sites. Disagreements between analysts were resolved by a third analyst revisiting the sample sites in question. In Region 2, approximately 30 percent of the sample sites were revisited by all three analysts to resolve interpretation differences.

Sampling Design

Given the choice of NAPP aerial photographs as the source of reference data, several probability sampling designs were considered and evaluated using the following criteria:

- Known inclusion probabilities, ensuring the objectivity of sample selection and the validity of statistical inferences;
- Small variance for estimated accuracy parameters;
- Good spatial distribution of the sample to ensure adequate precision for subregion estimates as well as precision of estimates for the full region;
- Representation of all classes, including rare classes;
- Low cost (both budgetary and time); and
- Simple to implement and analyze.

The key design element for controlling cost was to use NAPP photographs as primary sampling units (PSU) in a two-stage sampling design. This limited the number of photographs that had to be purchased, reduced the costs of photointerpretation, and lowered the potential cost of ground visits for confirming photointerpretation quality. Simple random or systematic sampling of pixels without this first-stage clustering structure would result in pixels dispersed among a much larger number of photographs. Similarly, a stratified random sample of pixels (strata identified by land-cover class) would also not have permitted control over the number of NAPP photographs sampled.

The second stage of the sampling design selected pixels as secondary sampling units (SSU) from the first-stage sample PSUs. The sampling design was then separated into two parts: a general, extensive design representing the full region and a special, separate design focusing on rare classes (Table 1). The general design was constructed so that all pixels, regardless of class, had an equal probability of being sampled. The special design for rare classes, on the other hand, was developed on the basis of stratification by rare land-cover classes to increase the sample size in these classes. The rare-class design focused on the objective of estimating the user's accuracy of the rare classes.

Sampling Frame

The sampling frame for the assessment consisted of the NAPP coverage for all of Federal Region 2. A gap in NAPP coverage (northwest part of the State of New York) resulted in the actual population assessed being smaller than the full region. Approximately 3 percent of the target region was not covered by NAPP photographs, and thus, the accuracy estimates apply to the remaining 97 percent of the region.

First-Stage Sample

To select PSUs for the objective of a spatially well-distributed sample, the entire sampling frame was partitioned into 333 grid cells on the basis of NAPP flight-line and frame numbers, with each grid cell measuring 15' by 15' and consisting of 32 NAPP photographs (four flight lines, eight photographs per line, Figure 1). Next, a stratified random sample was selected using the 333 grid cells as geographic strata (equal area for all strata). One photograph was selected at random from each grid cell, with all photographs having an equal probability of being

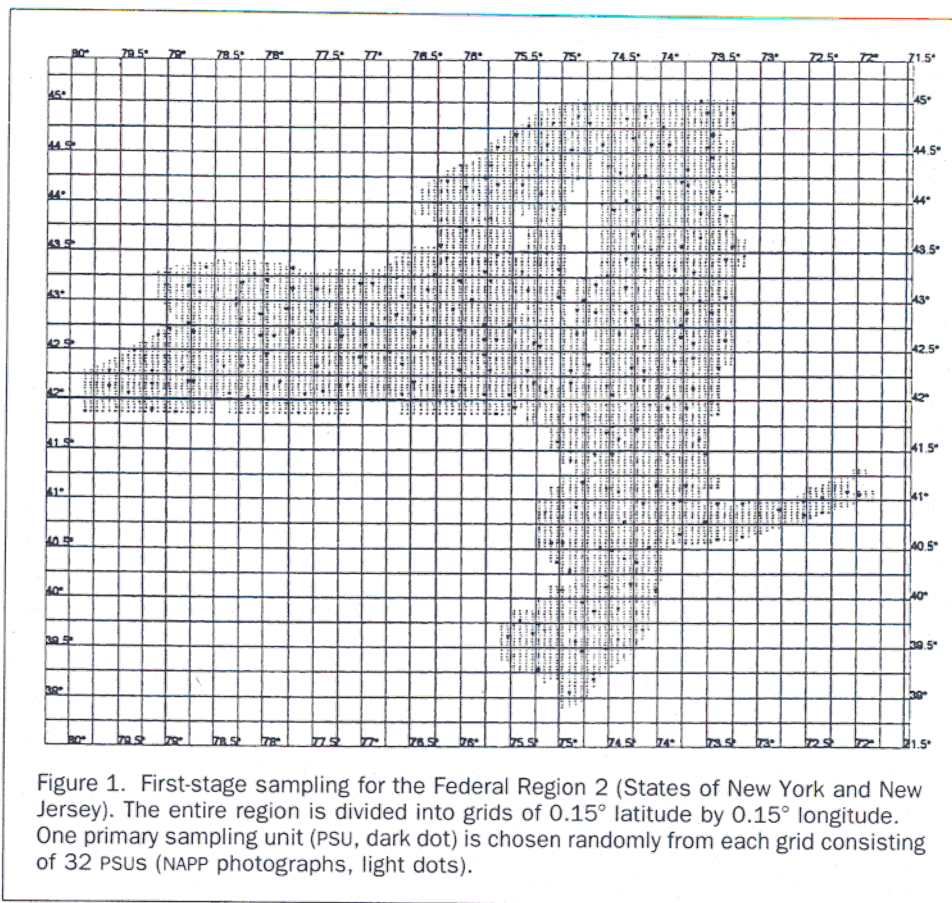


Figure 1. First-stage sampling for the Federal Region 2 (States of New York and New Jersey). The entire region is divided into grids of 0.15° latitude by 0.15° longitude. One primary sampling unit (PSU, dark dot) is chosen randomly from each grid consisting of 32 PSUs (NAPP photographs, light dots).

selected. It is important to note that the PSU is actually a "cropped" NAPP photo, not the full photo. Equal area regions (i.e., the interior of a NAPP photo) constitute the PSUs, and these first-stage sampling units are non-overlapping.

Grid cells on the boundary of Region 2 were treated as if they were complete strata (i.e., the grid cell or stratum contains 32 NAPP photographs). If the center of a selected NAPP photograph fell outside the regional boundary, it was not used in the sample. This restriction maintained the equal-probability characteristic of the first-stage design. If the boundary strata were not treated as each containing 32 cells, selecting one photograph from each cell would result in higher probabilities of sampling photographs along the region boundary. Such unequal probabilities are still allowable under a probability sampling protocol, but they create some extra complexity that was avoided in our analysis. The first-stage sample consisted of 278 NAPP photographs (Figure 1).

Second-Stage Sample

Second-stage sampling was accomplished by selecting four SSUs (pixels) within each PSU to provide the actual sample locations for obtaining the reference land-cover classification. Each photograph (PSU) was divided into equal-area quadrants, and one pixel was selected at random with equal probability from within each quadrant. Dividing the photograph into quadrants spatially distributed the sample pixels across the sampled photograph. If the PSU was a boundary photograph and the selected pixel was outside the target region, no sample pixel was obtained for that quadrant of the photograph. This design protocol extends the equal probability feature of the general design. A total of 1,033 SSUs were selected from the 278 NAPP photographs (Table 1).

Sample for Rare Classes

Seven land-cover classes were treated as rare classes, defined as classes each comprising less than 2 percent of the total map area and producing only a few SSUs from the general sample because of the equal-probability feature of this design (Table 1). An additional, separate design for rare classes was implemented for this study to augment the first, general design. The desire to exercise some control over the spatial distribution of the SSUs continued to be a key criterion influencing the rare-class treatment. Consequently, the NAPP photographs selected for the first-stage sample were used as the starting point for the rare class design. For each rare class, simple random sampling was used to select SSUs from all rare-class pixels found within the first-stage NAPP photographs. Within each rare-class stratum, pixels had equal inclusion probabilities, but these inclusion probabilities differed from those resulting from the general sampling design.

The sampling design described above produced the spatial distribution of sampled NAPP photographs shown in Figure 1 from which the second-stage samples of the general and rare-class designs were selected (Table 1). The first sample encompasses all mapped land cover-classes, whereas the second sample contains only additional SSUs for the rare mapped classes. These two samples can either be combined or treated separately for accuracy estimation.

Analysis and Estimation

The accuracy assessment results are derived from analysis of the error matrix summarization of the reference data (see Table 2 for error matrix notation). The equal-probability feature of the general sampling design permits using the conventional simple random sampling (SRS) formulas for overall accuracy (\hat{P}),

TABLE 2. SAMPLE ERROR MATRIX NOTATION.

		Reference				Sample Total	Population Total
		1	2	...	q		
Map Class	1	n_{11}	n_{12}	...	n_{1q}	n_{1+}	N_1
	2	n_{21}	n_{22}	...	n_{2q}	n_{2+}	N_2
			n_{k2}		n_{kq}	n_k	N_k
	q	n_{q1}	n_{q2}		n_{qq}	n_{q+}	N_{q+}
Column Total		n_{+1}	n_{+2}		n_{+q}	n	N

n_{ij} = number of pixels in map category i , reference category j

n_{k+} = number of pixels mapped as land-cover class k in the population

n_{+k} = number of pixels in land-cover class k in the sample according to the reference label

n = number of pixels in sample

N = number of pixels in population

user's accuracy for class i (\hat{P}_{ui}), and producer's accuracy for class j (\hat{P}_{Aj}). More efficient estimates of overall accuracy and producer's accuracy are available by using poststratification (Card, 1982). Poststratified estimators use the known pixel totals for each land-cover class (N_{i+}), treating the sample as a stratified random sample of n_{i+} pixels from the N_{i+} pixels in that class. Poststratification is justified by a conditional probability argument, in which the estimates are conditioned on the observed sample size in each stratum (Sarndal *et al.*, 1992, Section 7.10.2). Poststratification does not change the estimate of user's accuracy. For this study the overall accuracy (\hat{P}) and producer's accuracy (\hat{P}_{Aj}) are estimated using poststratified formulas, whereas user's accuracy (\hat{P}_{ui}) is based on the SRS formula: i.e.,

$$\hat{P} = \frac{1}{N} \sum_{k=1}^q \frac{N_{k+}}{n_{k+}} n_{kk} \quad (1)$$

$$\begin{aligned} \text{var}(\hat{P}) &= \frac{1}{nN} \sum_{k=1}^q N_{k+} \left(\frac{n_{kk}}{n_{k+}} \right) \left(1 - \frac{n_{kk}}{n_{k+}} \right) \\ &= \frac{1}{nN} \sum_{k=1}^q N_{k+} \hat{P}_{kk} (1 - \hat{P}_{kk}) \end{aligned} \quad (2)$$

(Equation 30, Card [1982])

$$\hat{P}_{ui} = n_{ii}/n_{i+} \quad (3)$$

$$\text{var}(\hat{P}_{ui}) = \hat{P}_{ui} (1 - \hat{P}_{ui}) / (n_{i+} - 1) \quad (4)$$

$$\hat{P}_j = \frac{(N_{j+}/n_{j+}) n_{jj}}{\sum_{k=1}^q (N_{k+}/n_{k+}) n_{jk}} \quad (5)$$

$$\begin{aligned} \text{var}(\hat{P}_{Aj}) &= \frac{N_{j+} \hat{P}_{jj}}{nNP_j^2} \left[\frac{N_{j+}}{N} \hat{P}_{jj} \sum_{k=1}^q \frac{N_{k+} n_{jk}}{N n_{k+}} \left(1 - \frac{n_{jk}}{n_{k+}} \right) \right. \\ &\quad \left. + (1 - \hat{P}_{jj}) \left(P_j - \frac{N_{j+}}{N} \hat{P}_{jj} \right)^2 \right] \end{aligned} \quad (6)$$

where $P_j = \sum_{k=1}^q \frac{N_{k+}}{N} \frac{n_{jk}}{n_{k+}}$ (Equation 28 of Card (1982)).

For the rare-class design, the first-stage sample is the same as that of the general-class design. Further, within a rare-class stratum, pixels are sampled with equal probability from all pixels of that class identified in the first-stage sample of PSUs. Consequently, the standard formulas for user's accuracy given by Equations 3 and 4 apply, with n_{ii} and n_{i+} being the sample values from the rare-class design. Because the rare-class design

excluded pixels from all "common" classes, producer's accuracy and overall accuracy are not estimated from the rare-class sample.

The variance estimation formulas represent approximations to the exact variance because the formulas assume that the general design is simple random sampling, and the rare-class design is stratified random sampling. Two design features are not accounted for by this assumption, the geographic stratification of the first-stage sample of NAPP photographs, and the clustering feature of the second-stage sample pixels. Not accounting for the geographic stratification tends to result in overestimating variance, whereas ignoring clustering structure generally results in underestimating variance. Neither potential source of bias in the variance estimators is likely to be large. The precision gained by geographic stratification is usually small, so a variance estimator not accounting for this slight decrease in variance will not be badly biased. Because only four pixels are sampled per cluster (NAPP photograph) in the general design, the effect of a high within-cluster correlation, which inflates the variance of cluster sampling, will also be small. Therefore, ignoring the variance inflation due to cluster sampling is not likely to result in a large underestimation of variance. The compensating effect of the two sources of bias (over and underestimation) further diminishes any bias concerns. We emphasize that no assumptions are needed to estimate the accuracy parameters themselves, and the SRS assumptions for the design apply only to variance estimation. The variance approximations used present considerable simplification of the formulas required to represent the full complexity of the two-stage sampling design.

Accuracy estimates are also obtained using only the high confidence sites (confidence index 3 or 4 in the response design). For Region 2, high confidence sites represent 82 percent of all sample sites. Recall that the probability sampling protocol permits no exclusions from the sample frame, so the reference sample may include pixels containing mixtures of land-cover classes as well as pixels intermediate between land-cover classes. The high confidence sites represent a statistical subpopulation of the full target population, so subpopulation estimation procedures are employed. The equal-probability feature of the sampling design makes the subpopulation analysis relatively simple. Suppose there are N' high confidence pixels in the Region 2 population. Given that n' high confidence sites appear in the sample, each high confidence pixel has a probability of n'/N' of being included in the sample. Although N' is unknown, it turns out not to be needed in the estimation formulas. An example will suffice to illustrate this. User's accuracy for land-cover class i for the high confidence sites is defined as $P'_{ui} = N'_{ii}/N'$, where N'_{ii} is the true number of high confidence pixels correctly classified as land-cover class i (P'_{ui} is the population parameter). The standard approach to estimate P'_{ui} is to estimate both N'_{ii} and N' ; thus,

$$\hat{P}'_{ui} = \hat{N}'_{ii}/\hat{N}' \quad \frac{(N'/n') n'_{ii}}{(N'/n') n'_{i+}} = \frac{n'_{ii}}{n'_{i+}} \quad (7)$$

and

$$\text{var}(\hat{P}'_{ui}) = \hat{P}'_{ui} (1 - \hat{P}'_{ui}) / (n'_{i+} - 1) \quad (8)$$

Note that N' does not appear in the estimate \hat{P}'_{ui} , nor does it appear in the estimated variance. By similar derivations, N' is not required for the estimates or variance estimates for producer's accuracy or overall accuracy when the estimation formulas are those of simple random sampling (or, in the case of the rare-class design, simple random sampling within strata). That is, the estimation formulas for the high confidence subpopulation

Class Number	Reference Classification														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Class Number	Reference Classification														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

[illegible]

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING

TABLE 5. ERROR MATRIX FOR THE GENERAL DESIGN USING ONLY HIGH CONFIDENCE SITES. ESTIMATED USER'S ACCURACIES AND STANDARD ERRORS ARE PRESENTED IN THE LAST TWO COLUMNS, AND PRODUCER'S ACCURACIES AND STANDARD ERRORS ARE IN THE LAST TWO ROWS. (ESTIMATED OVERALL ACCURACY IS 75.2 PERCENT WITH A STANDARD ERROR OF 1.5 PERCENT FOR THE HIGH CONFIDENCE SITES.)

Class Number	Reference Classification															n_{i+}	\hat{P}_{U_i}	$SE(\hat{P}_{U_i})$
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
1	77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	100	0.0
2	0	30	5	3	0	1	1	0	1	0	0	0	0	0	0	41	73.2	6.9
3	0	0	13	2	0	0	0	0	0	0	0	0	0	0	0	15	86.7	8.8
4	0	1	0	6	0	0	0	0	0	0	0	0	0	0	0	7	85.7	13.2
5	0	1	2	1	43	29	0	2	1	4	0	0	0	0	0	83	51.8	5.5
6	1	0	0	2	14	77	1	0	2	12	1	0	0	1	0	111	69.4	4.4
7	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	5	100	0.0
8	0	0	0	0	0	0	0	55	4	1	0	0	0	0	0	59	93.2	3.3
9	0	0	1	1	1	3	0	18	84	10	0	0	0	0	1	119	70.6	4.2
10	2	1	1	2	3	16	1	15	19	222	3	2	0	0	1	288	77.1	2.5
11	0	0	0	0	0	0	0	6	2	3	6	0	0	0	0	17	35.3	11.6
12	1	0	0	0	0	0	0	0	0	0	1	8	0	0	0	10	80.0	12.7
13	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0.0	0.0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	100	0.0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	—
n_{+j}	81	33	22	18	61	126	8	96	112	252	11	10	0	2				
\hat{P}_{+j}	97.5	89.1	51.2	41.0	68.9	62.5	64.5	54.6	78.8	86.6	61.7	75.2	0.0	34.2				
$SE(\hat{P}_{+j})$																		

TABLE 6. USER'S ACCURACIES FOR RARE-CLASS SAMPLE USING ONLY HIGH CONFIDENCE SITES.

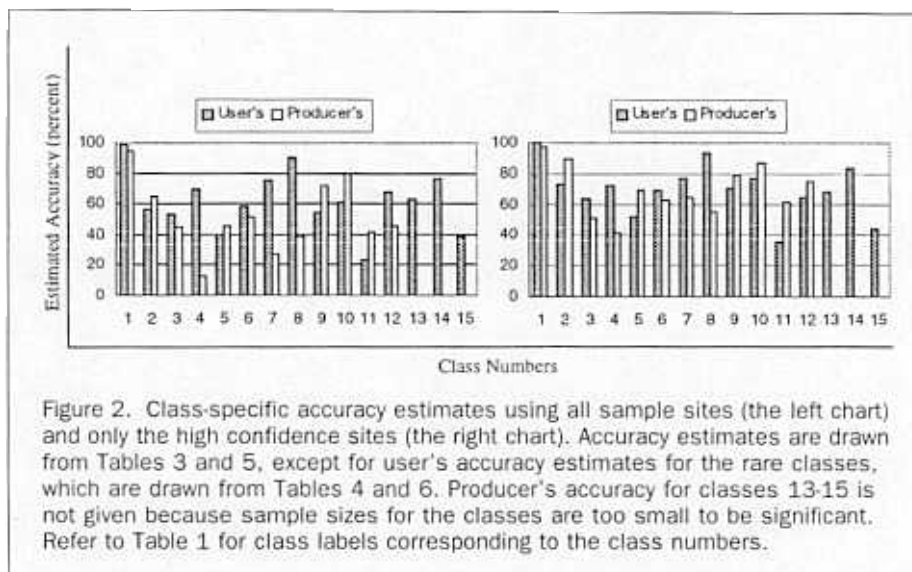
Class number	n'_{i+}	Proportion in all sample sites	\hat{P}_{U_i}	$SE(\hat{P}_{U_i})$
		0.77	63.6	
		0.83	71.8	
		0.93	76.9	
12		0.83	64.1	
13		0.93	68.4	
14		0.87	83.9	
15		0.87	44.1	

Landsat classification may have correctly classified a pixel as hay/pasture, but the NAPP photograph flown a year later may show that the land has been rotated to row crop. This type of mismatch was handled in two ways. First, if the tonal difference between TM and NAPP could be unambiguously determined, then the land-cover label would be based on knowledge

of vegetation phenology. Otherwise, a rating of low confidence would be given as a practical way to minimize such non-misclassification error differences. This effect can be seen from improved class accuracies in Table 5 when the analysis is of the high confidence subpopulation for these two land-cover types.

The accuracy of the forest wetland class (2.52 percent of the mapped area) is low. This land-cover class was derived in part from multiple data sources, including New York State Regulatory Wetlands data, New Jersey State land-cover data, and the 1970s USGS land-use and land-cover data. These data sets were developed at different time periods and for different purposes, and they are not ideal for regional consistency, temporal consistency, or level of detail. Additionally, depending on dates of the image data and NAPP photographs, the presence and optical properties of ground water can affect whether the land cover is classified as woody wetland or forest, as suggested by confusions in Tables 3 and 5.

The class of transitional barren (0.10 percent of the map) may also suffer the same deficiency related to timing and/or the interpretation capabilities of the two primary data



involved: TM and NAPP. The class is designed for conditions such as temporary clearing and regeneration of forest cover. Because of the inevitable date difference between TM and NAPP, it is possible that what is considered transitional barren at the TM date may already have enough vegetation to be called, say, young conifer stand at the NAPP date (see Table 4).

Low accuracy for classes that are land use in nature is understandable. Despite the extensive use of ancillary data, such as population census, it is very difficult to unambiguously separate high intensity residential from urban use, either during the modeling of TM data or simply when viewing it on a NAPP photograph. The same is true for the artificially designated barren classes between quarry/strip-mine class and sandy/gravel class. If the 15 classes were aggregated to Anderson level 1 (Anderson *et al.*, 1976), the estimated overall accuracy using all sample sites would be improved from 63 percent to 80 percent, an indication that a substantial amount of confusion is due to similarly defined classes.

Summarizing the above analyses of performances of individual classes makes it clear that land-cover mapping accuracy is strongly related to homogeneity of the land use. An examination of the spatial distribution of misclassification errors (Plate 1) shows that most misclassification errors in the land-cover map occur at heterogeneous fringes of multiple land-cover and land-use patterns. For example, the extensive forest cover on the Adirondack Mountains in northeast New York is relatively error free. Except for differences owing to timing between the TM and NAPP data sources, which are arguably not misclassification error, narrow definitions in land cover (e.g., mixed versus conifer forests) or land use (e.g., different types of barren land) seem to be the primary causes of misclassifications.

Photointerpretation

NAPP aerial photographs provided the best available reference material under the constraints of the USGS regional land-cover mapping program. Interpretation of 1:40,000-scale aerial photographs is a feasible and practical way to collect reference data for the regional accuracy assessment. Visually locating sample sites on the photographs takes time, but the precision is generally satisfactory. There are two drawbacks to using this approach: (1) the often-unavoidable time differences between the TM and NAPP dates, as discussed above, and (2) the need for field visits to ascertain land cover for low confidence sites.

About 18 percent of the sample sites are low confidence sites. It is important to note that low confidence sites are not necessarily related to mixed land-cover classes. Rather, low confidence is recorded often because the interpreter feels that the land cover is simply too difficult to read on the NAPP photograph. This may be due to an edge condition, such as between water and land, or due to lack of information on land use (e.g., high intensity residential versus commercial use). The interpreter could use his own knowledge or other features available on the photograph to infer the land cover, but doing so would often lead to a low confidence rating. Interpretation of low confidence sites may be improved by field visits to these sites. In this study, field visits were not conducted due to the limited time available to the project staff.

The fact that accuracy increases sharply for the high confidence sites tells us that limiting accuracy sampling to clearly interpretable (homogeneous) pixels would have provided a much more optimistic view of accuracy. A conventional way of making the map more homogeneous is by limiting sample sites to only homogeneous pixel blocks (e.g., a window of 3 by 3 pixels). In the case of the USGS regional land-cover mapping, no filtering is used to smooth the resulting land-cover maps, so pixel blocks were not used for accuracy assessment.

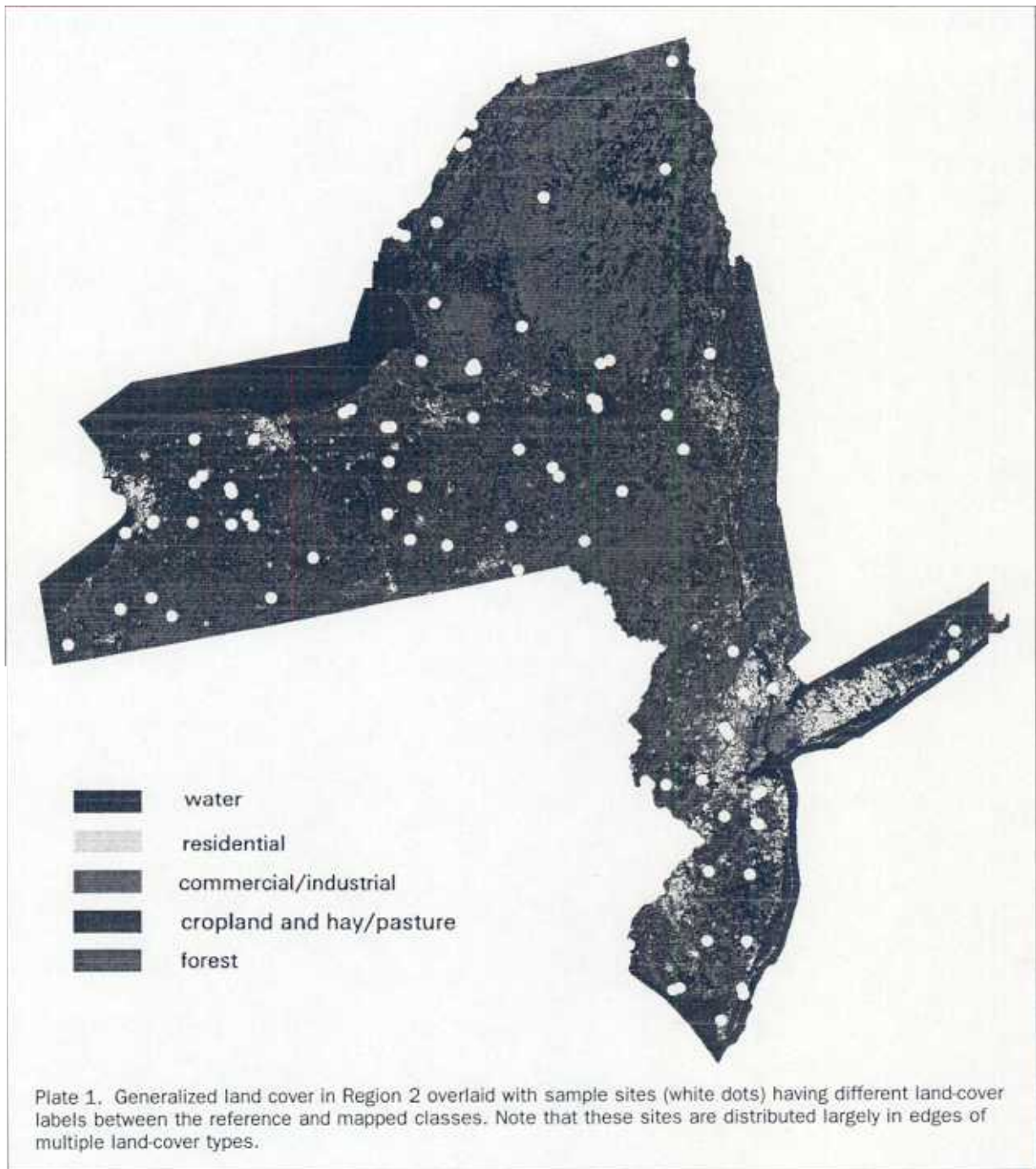
Although it is "difficult to examine closely the source and nature of errors in classifications using hard [classification] methods" (Zhang and Foody, 1998, p. 2722), the data collected

in our response design protocol provide some insight into the potential errors of the reference classification and the impact of these errors on the reported accuracy. Recording a secondary land-cover class, if one existed, for the reference sample pixel allowed for employing a "softer" measure of agreement and subsequent analysis akin to the approach (using fuzzy classification methods) suggested by Zhang and Foody (1998, p. 2726). The confidence index data allowed for analyses of error by subsets of the data. For example, results from the high confidence (confidence of 3 or 4) subpopulation may be considered as representative of the accuracy of "pure" pixels. Comparing the accuracy estimates for the high confidence pixels to the estimates for all sites provides an indication of the contribution to classification error attributable to the sources of a low confidence index (e.g., homogeneous pixel but land-cover type is not clearly interpretable as of the defined MRLC classes, mixed pixel containing two or more land-cover classes, changes in land cover between the date of imagery and the date of the NAPP photo, and photointerpreter disagreement). Joria and Jorgenson (1996, p. 167) employed a similar analysis in which they assigned a qualitative confidence level from 1 to 3 (1 being most confident) and reporting accuracy for the different confidence levels. Zhang and Foody (1998, Table 1) conducted an analogous analysis by reporting overall accuracy for the subpopulation of pure pixels which can be compared to accuracy achieved for subpopulations including non-pure pixels. The response design protocol and additional analyses implemented in Region 2 address some of the issues affecting the "conservative bias" of accuracy (Verbyla and Hammond, 1995) resulting from an assessment in which the sample is not restricted to homogeneous land-cover areas or pure pixels.

Discussion of Sampling Design Issues

The two-stage, cluster sampling design employed for both the general and rare-class designs is consistent with the approach taken by Belward *et al.* (1996), Edwards *et al.* (1998), and Lillesand (1994) for large-area accuracy assessments. Several advantages derive from our chosen designs. The general design is equal probability, but not SRS. It retains much of the ease of analysis of an SRS and exercises strong control over the spatial dispersion of the sample. A similar characteristic holds for the rare-class design. The rare-class design is stratified (but not stratified random) at the first stage, leading to equal inclusion probabilities within each stratum. The within-stratum design is not SRS, however. This structure is particularly advantageous for ground-visited reference data. Edwards *et al.* (1998) noted that randomly distributing the sample points within strata would have severely compromised their attempt to maximize logistical efficiency for sampling the large area represented by the Utah GAP land-cover map. The two-stage clustering structure employed in their study, and selected for the Region 2 assessment, alleviates that obstacle.

Aronoff (1982), Rosenfield *et al.* (1982), Congalton (1991), and Edwards *et al.* (1998) mention the possibility of combining a stratified sampling design with an equal probability design. The usual motivation for employing two designs is that the equal probability design (typically a simple random or systematic sample) can be quickly and easily implemented, and it does not require the land-cover map to be complete, thus allowing the reference sample data to be collected at the same time as the imagery is taken. However, an equal probability design such as simple random or systematic sampling will result in small sample sizes for the rare classes unless the overall sample size is extremely large. The allocation of sample sizes to land-cover classes resulting from our general sampling design (Table 1) illustrates the inadequate coverage for precise estimation of rare-class accuracies. The second design, stratified by mapped land-cover types, ensures representation of the rare cover types.



For Region 2, the data from just the rare-class design provided the information to estimate user's accuracy with reasonably good precision for the rare classes. The advantage of combining the general and rare-class samples accrues to estimating producer's accuracy for rare classes. The theory for combining probability samples exists (Hartley, 1974; Sarndal *et al.*, 1992, p. 545). The estimation formulas require fairly elaborate data management procedures ("bookkeeping"), and the variance estimators can be complex when the two sampling designs include strata and two-stage sampling. Consequently, these formulas and the resulting estimates are not presented in this paper, with the implication on our analyses being that producer's accuracies for some of the rare classes can be estimated with better precision than that shown by the estimates presented for just the data from the general sampling design.

Our initial motivation for employing both designs, rather than just one design stratified by all land-cover classes, was the expected advantage of enhanced flexibility for analyses by users of the accuracy data. That is, we anticipated that users would subject the land-cover map to diverse applications, and that various aspects of accuracy would be of interest to different users depending on their application. For example, some users might be interested in specific subregions of the map, whereas others may be interested in aggregating certain land-cover classes. The equal probability feature of the general design facilitates ease of such analyses because the weighted analysis required of a stratified sampling design would not be necessary for the general design. The general design was expected to be less precise than a stratified design for estimating user's accuracy of rare land-cover classes. But the poststratified

analysis of the general design should result in precision similar to a proportionally allocated stratified design for overall and producer's accuracies, so the precision disadvantage of the general design was expected to be small for these estimates.

In hindsight, these advantages of the general design are probably not compelling relative to its disadvantages, and our decision to employ the general design may reflect an oversensitivity to the needs of potential secondary analyses based on the MRLC accuracy data. A viable, practical alternative to the two-design approach we used would be to base the entire assessment on a single design stratified by all land-cover types. This design would still retain the two-stage cluster sampling protocol, and it would be equal probability within strata, thus retaining some of the simplicity of analysis gained by our general design. A single stratified design would avoid the complexity of the dual-frame estimation methods required for analyzing the combined general and rare-class designs, and it would have allowed for a more balanced allocation of sampling effort among the common and rare classes. For example, the general design of Region 2 resulted in 370 sample sites for broadleaf deciduous forest. Consequently, accuracy of this class was estimated very precisely. However, a more efficient use of sampling resources may have been to allocate some of the deciduous forest pixels to other classes in order to improve the precision of the estimates for those classes. Having a single design stratified by land-cover class would have permitted achieving this more equitable allocation. Lastly, because the reference data were photointerpreted, one of the other advantages of a two-design approach — sampling for reference data simultaneous to the time of the imagery — was not relevant.

Whether a stratified design is employed for all land-cover classes or for just the rare classes, several features of the sampling design merit further evaluation. The main issue is allocation of samples to PSUs. A potential disadvantage of the two-stage cluster sample occurs if a land-cover class is spatially clustered within in a small region, resulting in most of the sample pixels of this class being found in a few PSUs. In a worst-case scenario, all the sample pixels could be in a single first-stage PSU. Although the design is still a probability sampling design, the precision of the estimates for this spatially clustered class will likely be poor because of the strong clustering feature of the design. To circumvent this feature of cluster sampling, we would like the sample to be distributed among a larger number of PSUs. Two options are proposed. The Region 2 design selected the second-stage sample with equal probability from all pixels of the rare class identified in the first stage sample. Consequently, the second-stage pixels will be represented in the PSUs in proportion to the number of pixels of that class in the PSU; i.e., PSUs with many pixels of the class will have more of the second-stage sample pixels. The second-stage design can be changed so that, for example, a single pixel could be sampled from each PSU. This would effectively spread the second-stage sample for this land-cover class among a larger number of PSUs, diminishing the precision disadvantage of clustering. However, the consequence is that the second-stage sample now has a more complicated to analyze unequal probability sampling scheme, and it is possible that this will also create higher variances for the estimates.

A related dimension of this clustering problem occurs when a rare class only appears in one or two first-stage sample PSUs. A design modification would therefore need to focus on how to increase the representation of rare classes in the first-stage sample of PSUs. A typical trade-off is involved. PSUs having rare-class pixels are identifiable from the land-cover map, so it would be possible to sample such PSUs with higher probability. The added complexity of the unequal probability structure would again need to be dealt with, and it is not clear how this change would affect precision of the estimates for the non-

rare classes. The design protocol must also accommodate the existence of several rare classes, and this could further complicate a first-stage protocol designed to sample PSUs with multiple rare classes with higher probability. In Region 2, clustering of sample pixels into one or two PSUs did not occur, but this potentiality motivates investigation of design structures to prevent this problem from occurring. A practical problem to implementing this change in the first-stage sampling protocol is that the land-cover composition for each NAPP photo within the entire region must be described. The cost of such a preliminary analysis may be prohibitive.

Conclusions

The USGS regional land-cover mapping program is conducted over very large areas with a relatively large number of land-cover and land-use classes. Some of the land-use classes (such as the three barren classes, and residential and commercial classes) are very similar spectrally, posing a challenge to both the mapping and photointerpretation during accuracy assessment. Given these conditions, the overall and class-specific accuracy estimates for Region 2 are generally satisfactory. Most misclassification errors occur along edges of heterogeneous land-cover and land-use patterns, and a majority of the confusion is between related land-cover or land-use classes.

Adherence of the Region 2 sampling design to probability sampling protocol resulted in a statistically defensible accuracy assessment, and the estimates apply to a well-defined population, the 97 percent of the area of Region 2 for which NAPP photography was available. The accuracy estimates are statistically consistent, as recommended by Stehman and Czaplewski (1998). The two-stage cluster sampling feature of both the general and rare-class designs results in an equal probability sample (for all classes for the general design, and within strata for the rare-class design), facilitating ease of analysis. This design also provided the advantage of a spatially well-distributed sample across Region 2, yet at the same time created logistical efficiency by restricting the sample spatially to the area within the first-stage sample NAPP photos. Combining the general and rare-class designs to improve precision of producer's accuracy for the rare classes created a more complicated analysis protocol than we had envisioned. In retrospect, the advantages of employing two separate designs were probably not sufficiently strong to merit this approach over some simpler alternatives. In particular, we propose retaining the two-stage cluster sampling feature for subsequent assessment of other EPA Federal regions, but we recommend using a single design stratified by all land-cover classes, not just the rare classes. When combining data from different strata, the analysis of this single design must incorporate the appropriate strata weights required for consistent estimation of accuracy measures, but such stratified sampling analyses should be routine in accuracy assessment work. Employing a different sampling design from the Region 2 design in other EPA federal regions will not adversely affect an eventual summary of accuracy at the conterminous U.S. level. In the design for the entire U.S., the federal regions represent strata and, as such, each may have its own separate design. A complete summary for all ten federal regions will be reported when the accuracy assessment results are complete.

NAPP photographs provided a practical and economic means for assessing large-area land-cover accuracy in the United States. Visually locating sample sites and interpreting land cover proved effective and efficient. There are, however, some weaknesses in this practice, chiefly the associated uncertainties of land-cover change owing to the difference in the TM and NAPP acquisition dates, and the difficulty in land-use interpretation. The confidence index used in Region 2 is one way to evaluate these uncertainties. More effective but not overly complicated measures may be needed to better address such problems in future operational large-area accuracy assessment.

Acknowledgment

The work reported on herein was performed under U.S. Geological Survey contract 1434-CR-98-CN-40274.

References

- Anderson, J.R., E.E. Hardy, J.T. Roach, and R.E. Witmer, 1976. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*, U.S. Geol. Survey Prof. Paper 964, U.S. Geological Survey, Washington, D.C., 28 p.
- Aronoff, S., 1982. The map accuracy report: A user's view, *Photogrammetric Engineering & Remote Sensing*, 48:1309–1312.
- Belward, A.S. (editor), 1996. *The IGBP-DIS Global 1Km Land Cover Data Set (DISCOVER): Proposal and Implementation Plans*, IGBP-DIS Working Paper #13, Joint Research Centre, Space Applications Institute, Ispra, Italy, 66 p.
- Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy, *Photogrammetric Engineering & Remote Sensing*, 48(3):431–439.
- Cochran, W.G., 1977. *Sampling Techniques, Third Edition*, Wiley, New York, 428 p.
- Congalton, R.G., 1991. A review of assessing the accuracy of classification of remote sensed data, *Remote Sensing of Environment*, 37:35–46.
- Edwards, T.C., Jr., G.G. Moisen, and D.R. Cutler, 1998. Assessing map accuracy in a remotely-sensed ecoregion-scale cover map, *Remote Sensing of Environment*, 63:73–83.
- Hartley, H.O., 1974. Multiple frame methodologies and selected applications, *Sankhya Series B*, 36:99–118.
- Joria, P.E., and J.C. Jorgenson, 1996. Comparison of three methods for mapping tundra with Landsat digital data, *Photogrammetric Engineering & Remote Sensing*, 62:163–169.
- Lillesand, T.M., 1994. Strategies for improving the accuracy and specificity of large-area, satellite-based land cover inventories, *Proceedings, ISPRS Mapping and GIS Symposium*, 31 May–03 June, Athens, Georgia, 30:23–30.
- Rosenfield, G.H., K. Fitzpatrick-Lins, and H.S. Ling, 1982. Sampling for thematic map accuracy testing, *Photogrammetric Engineering & Remote Sensing*, 48:131–137.
- Sarndal, C.E., B. Swensson, and J. Wretman, 1992. *Model Assisted Survey Sampling*, Springer-Verlag, New York, 694 p.
- Stehman, S.V., and R.L. Czaplewski, 1998. Design and analysis for thematic map accuracy assessment: fundamental principles, *Remote Sensing of Environment*, 64:331–344.
- Verbyla, D.L., and T.O. Hammond, 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids, *International Journal of Remote Sensing*, 16:581–587.
- Vogelmann, J.E., T.L. Sohl, P.E. Campbell, and D.M. Shaw, 1998. Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources, *Environmental Monitoring and Assessment*, 51:415–428.
- Zhang, J., and G.M. Foody, 1998. A fuzzy classification of sub-urban land cover from remotely sensed imagery, *International Journal of Remote Sensing*, 19:2721–2738.

(Received 04 June 1999; revised and accepted 15 November 1999)

Photogrammetry

Remote Sensing

LIDAR

Radar

GIS

GPS

**Biological
applications**

**Geological
applications**

CALL FOR PAPERS

PE&RS Special Issue on Geospatial Information Technology in KOREA

In December 2001, the American Society for Photogrammetry and Remote Sensing will devote its issue of *Photogrammetric Engineering and Remote Sensing (PE&RS)* to Geospatial Information Technology in KOREA.

Authors are encouraged to submit manuscripts on "cutting edge" research underway in KOREA in the following areas

- Photogrammetry ● Remote Sensing ● LIDAR
- Radar ● GIS ● GPS ● Biological & geological applications
- All aspects of geospatial information technologies

All manuscripts must be prepared according to the "Instructions to Authors" published in each issue of *PE&RS* and on the ASPRS web site at www.asprs.org. Papers will be peer-reviewed in accordance with established ASPRS policy. **Manuscripts must be received by March 31, 2001 to be considered for publication.**

Please send completed manuscripts or direct inquiries to :

Dr. Woosug Cho, Guest Editor
Department of Civil Engineering
Inha University
253 Yonghyun-Dong Nam-Gu
Inchon, 402-751, KOREA
FAX : +82-32-873-7560
wcho@inha.ac.kr